Ins. A1

PROCESS FOR RELAYING IP APPLICATION FRAMES WITHIN AN ATM NETWORK SWITCH WITH DISTRIBUTED ARCHITECTURE.

Ins. A2

The present invention relates to a process for relaying IP application frames within an ATM network switch with distributed architecture and egress storage.

Ins. A3

Communications networks known by the abbreviation ATM standing for "Asynchronous Transfer Mode" allow the dissemination of fixed-length packets dubbed "ATM cells", consisting of a five-byte header and a forty-eight-byte payload. The header contains in particular a logical path identifier, dubbed the VPI/VCI field, the abbreviation standing for "Virtual Path Identification and Virtual Channel Identifier" which allows the cell to be steered through the switches which it encounters on its journey between the sending user and the destination user.

The applications which are able to use ATM networks for communicating the data are very diverse. Most of the applications capable of using ATM networks have a format specific to their data: it may, for example, involve frames in the IP format of the "INTERNET" protocol or else frames in the format of the MPEG standard where MPEG is the abbreviation standing for "Moving Picture Expert Group". Adaptation between the format of the application frames and the format of the ATM cells is performed in a layer known as the adaptation layer, designated by the abbreviation AAL standing for "ATM Adaptation Layer", this layer being in particular responsible for segmenting the frames into cells and conversely for reassembling the cells received from the network into frames.

Any ATM switch employs, in the manner represented in Figure 1a, four major sets of functions, an access function 1 for accessing each port of an ATM switch, an ATM layer function 2, a cross-connection function 3 and a management function 4.

The access function 1 ensures the conversion of the ATM cells into the format suitable for the transmission carrier linked to the said port and vice versa. This function makes it possible to present the inbound cells to the ATM layer in a single format which is independent of the nominal throughput and of the optical, electrical, radio, etc. technology of the transmission carrier from which they originate. The ports of a switch make it possible to link several switches together but they also make it possible to link a user of the ATM services to a switch.

The processing operations to be implemented in the access function are described in an ample standards literature, both from the ANSI and from the ITU and the ATM Forum. The major classes of interface which are defined in these documents are:

The PDH interface, the abbreviation standing for "Plesiochronous Digital Hierarchy", defined in the document ITU-T G.804, G.703.

The SDH interface, the abbreviation standing for "Synchronous Digital Hierarchy", defined in the document ITU-T G.708, etc.

The SONET interface, the abbreviation standing for "Synchronous Optical Network", defined in the document ANSI-T 1.105, etc.

The 25.6 Mbit/s IBM interface defined in the document af-phy-0040.000.

The ATM layer function 2 groups together several functions such as in particular the management of the cell headers, the translation of the VPI/VCI logical paths, the abbreviation standing for "Virtual Path Identification and Virtual Channel Identifier", the processing of OAM management cells, the abbreviation standing for "Operations Administration and Maintenance", an important part of the traffic management which comprises the subfunctions known by the abbreviations UPC standing for "Usage Parameter Control", SCD standing for "Selective Cell Discard", EPD standing for "Early PDU Discard", RM cells the abbreviation standing for "Resource Management", etc.

The processing operations to be implemented in the ATM layer function are described in particular in the following standards documents from the ITU and the ATM Forum:

- B-ISDN ATM Layer Specification [ITU-T I.361]
- B-ISDN Operation and Maintenance Principles and Functions [ITU-T I.610]
- Traffic Management Specification Version 4.0 [AF-TM 4.0]

The cross-connection function 3 switches the cells from an ingress direction to one or more egress directions, depending on indications formulated by the ATM layer when translating logical paths.

This function is at the heart of any ATM switch and it has formed the subject of an ample literature which need not be recalled here. The cross-connection ring and the cross-connection network constitute two frequent types of implementation of this function.

The management function 4 comprises subfunctions such as: local supervision of the switch (alarms, discovery of the configuration of the switch and of the local topology, management of versions, etc.), dialogue with the centralized supervision of the network, the dialogues required for establishing switched virtual circuits, etc.).

For a more detailed description of some of these subfunctions, reference may be made for example to the standards literature of the ATM Forum:

- ATM User-Network Interface (UNI) Signaling Specification Version 4.0 (af-sig-0061.000)
- Private Network-Network Interface Specification Version 1.0 (af-pnni-0055.000)
- Integrated Layer Management Interface (af-ilmi-0065.000)

These various functions interface with one another as indicated below. It should be noted that the management function behaves exactly like a user except that its linkup with the ATM layer does not pass through an external port of the switch, and, therefore, does not require any access function. On the other hand, the management function deals only with ATM cells and also with messages which it must therefore segment and reassemble by way of an AAL adaptation layer which therefore constitutes an additional function: the adaptation function.

A conventional switch architecture consists in distributing the functions over distinct hardware facilities, possibly duplicated so as to allow backup of a defective facility of the same nature, and which are implanted in the switch in sufficient number to satisfy the processing load forecast as a function of the configuration of the network at this spot. In practice, these facilities are electronic-component cards joined together in a rack and conversing with one another via one or more data buses fitted to the backplane. They define what is commonly referred to as a "distributed architecture".

Traditionally, the hardware architecture of a distributed ATM switch distinguishes, as shown in Figure 1b, three types of modules: a cross-connector module 5, a management module 6 and junctor modules 7_1 ... 7_n. The functions of the switch are shared among these various modules with the constraint however that the junctor modules deal at least with the access function, the cross-connector module 5 with the cross-connection function and the management module 6 with the management function.

In Figure 1b, the links 8_1 ... 8_n existing between each junctor module and the cross-connector module are called "cross-connector junctions". Furthermore, each junctor implements an access function capable of managing one or more ports. When a cell passes through a switch, it begins by passing through a first junctor, the so-called "ingress junctor" for this cell, then a second junctor, the so-called "egress junctor". Since several ingress junctors can simultaneously despatch cells to one and the same egress junctor, cell congestion may occur on account of the limited egress throughput of this egress junctor. Cell storage and queuing mechanisms are then triggered while waiting for the congestion to be cleared. These storage mechanisms may be located at the ingress, at the egress, in the cross-connector or in several of these elements at once. One then speaks of an architecture with "ingress storage", "egress storage", etc.

The users of a communications network can envisage several modes for exchanging their data. These modes are represented schematically in Figures 2a to 2f. The point-to-point mode, Figure 2a, puts in touch two users A, D exclusively, each of them being able to be a sender and receiver. In this mode, anything which one of the users sends is received by the other. A variant of the point-to-point mode consists in specializing the sender or receiver roles of each of the two users (unidirectional point-to-point communication).

The point-to-multipoint mode, Figure 2b, puts in touch more than two users A, C, D, one of whom is exclusively a sender and the others exclusively receivers. Anything which is sent by the sender is received by all the receivers.

The multipoint-to-point mode, Figure 2c, also puts in touch more than two users A, B, C, one of whom is exclusively a receiver and the others exclusively senders. Anything which is sent by one of the senders is received by the receiver.

Finally, the multipoint-to-multipoint mode, not represented, puts in touch at least two [sic] users A, B, C, D, each able to be a sender and receiver. In this last mode, anything which is sent by any one of the users is received by all the other users and also by the sender.

The multipoint-to-multipoint and point-to-multipoint communications are especially natural in the case of a shared-medium communication network such as Ethernet networks as represented schematically in Figure 2e. Indeed, in this case, all the users are linked to a single medium and all the stations A, B, C, D connected to this medium receive all messages despatched by the other stations. On the contrary, in the case of an ATM network, as represented schematically in Figure 2f, the broadcasting to multiple destinations A, B, C, D of a cell sent by one of the users requires that the network should itself generate the copies of the cell in question.

The term "connection" refers to any communication according to one of the modes defined hereinabove, between a well-defined set of users, this communication being endowed with a specific list of attributes such as: service quality parameters, traffic parameters, etc.

The implementation within an ATM network of communications in the various modes defined hereinabove may be considered from several points of view, in particular: signalling, routing, conveying of the data and management of the resources.

As far as point-to-point connections are concerned, the signalling and routing aspects are amply described in the documents ([ITU-T Q.2931], [AF-SIG 4.0], [AF-PNNI1.0], [AF-IISP]) of the standards literature.

They consist in determining a route through the network between the two users, such that this route satisfies the traffic and service quality constraints of the connection. The route is characterized by a list of highways. Each switch of the route allocates the connection a logical path number relating to the ingress highway of the connection into the switch and maintains a translation table which matches up this identifier with the outbound direction to be taken by the cell and the logical path identifier of the connection in the next switch. Thus, any cell of a connection can be steered gradually simply by consulting the logical path identifier present in the cell header and the local translation table.

Sub B1

In a switch with distributed architecture, such as the one of Figure 1b, this translation can be performed by the ATM layer function of the ingress junctor. The cell is then sent back to the cross-connector module together with an indication of the egress cross-connection junction to which the cross-connector must switch the cell. This indication can be despatched in a specific header prefixed to the start of the cell. Translation devices in accordance with this particular case have been described by the applicant, for example in French patent applications No. 2 670 972, 2 681 164, 2 726 669 and French application FR 97 07355 not yet published.

Symbolically, the point-to-multipoint connections can be represented by a "tree" with a "root" representing the sender user and its "leaves" representing the receiver users. The implementation of this type of connection is standardized as regards the signalling and routing. It involves simply forming a point-to-multipoint connection by first creating a point-to-point connection and then by grafting new leaves onto it. This adding of leaves can be done on the initiative of the root or else of the leaf.

Sub B2

As regards the conveying of the cells, the model of the point-to-point connection, that is to say the ingress translation only, cannot always be applied. In Figure 3a where the elements counterpart to those of Figure 1b are labelled with the same references, a point-to-multipoint connection is represented. This connection enters the switch via a port P1 and leaves it via the ports P2, P4, P5, P7. In this case, the ingress translation envisaged above can order the cross-connector 5 to copy each cell of this connection to the three cross-connector junctions concerned (7_3, 7_4, 7_n) but it is not capable of indicating the egress ports to which the cell should be despatched. To do this, it is necessary to append this information to the translation tables and to convey it from the ingress to the egress. Moreover, the egress logical path depends on each egress port and it cannot be allocated a unique logical path for all the egresses.

For all these reasons, it is generally preferred to carry out a double translation: an ingress translation which replaces the logical path of the cell by a "broadcast index" representative of the connection within the switch, then a second translation at egress which translates the broadcast index into a list of pairs of the form (port, logical path) together with any necessary copies.

The multipoint-to-point and multipoint-to-multipoint connections are not at present dealt with in the standardizing facilities concerned with ATM. Hence, there is for the moment no signalling nor routing defined for this type of connection.

In terms of cell conveying, any communication topology which brings data from different geographical origins to converge to one and the same link poses the problem of so-called "interleaving of the PDU application frames", PDU being the abbreviation standing for "Protocol Data Unit". Indeed, the application frames (PDUs) being segmented by the AAL layer into cells, the cells of various frames arrive interleaved at the destination. To reassemble the frames, the destination would have to be able to rediscover which frame each cell belongs to. Now, the segmentation mechanism used most commonly in UBR connections, implemented within the AAL adaptation layer 5, does not allow this identification. It allows only the identification of the last cell of the PDU frame, this being sufficient in point-to-point or point-to-multipoint modes since the ATM cells are transmitted in sequence.

Despite all these problems, communications needs in multipoint-to-point and multipoint-to-multipoint mode exist. They could be dealt with theoretically by superposing point-to-point or point-to-multipoint connections. The case is apparent in particular within the framework of local area network emulation known by the abbreviations LAN Emulation or LANE, standing for "Local Array Network Emulation", where any sophisticated mechanism is set in place for emulating a shared medium (ELAN) within an ATM network [AF LANE]. Each user is assigned an LEC function (Lan Emulation Client). The shared-medium emulator makes it possible to achieve multipoint-to-multipoint communications by using the server known by the abbreviation BUS, standing for "Broadcast or Unknown Server" which is defined in the LANE standard, and whose architecture is shown in Figure 3b, which a user can employ to transmit messages broadcast to all the users of an emulated shared medium, or to another user to whom he is not yet directly linked. Each user of the ELAN possesses a point-to-point connection to the BUS server and the BUS server possesses a point-to-multipoint connection to all the users of the ELAN, as indicated in Figure 3b.

Another example of a need in multipoint-to-point and multipoint-to-multipoint communications is provided by the emulation of routing between local area networks. This function can in particular be implemented according to the standard known by the abbreviation MPOA, standing for "Multiprotocol Over ATM" of the ATM Forum which makes it possible to perform a virtual routing between various emulated local area networks (ELANs) or various virtual local area networks (VLANs) or an ELAN [AF MPOA]. Another way of performing the routing emulation consists in embedding routing software within the management unit of the ATM switch. Such an embedded routing emulation function is designated hereinafter as virtual router. In this context, the virtual router is akin to a user of the various ELANs which it interconnects. In this regard, an LEC function (router LEC) must correspond thereto for each ELAN. The virtual router must be implemented in a switch, for example by a specific procedure executed within the management module, thereby incurring the risk of clogging up the said module when the exchanges between the ELANs are sufficiently supported.

Ins B1

The aim of the invention is to alleviate the aforesaid drawbacks.

To this end, the subject of the invention is a process for relaying IP frames in the form of PDU application frames within an ATM switch with distributed architecture and egress storage comprising a management module and several ingress and egress junctors having a routing emulation function ensuring IP frame routing between the users of various ELAN media and represented in each of these ELANs by its router LEC module, characterized in that it consists in offloading the frame relay function into the ATM layer of the junctors by examining the first cell of each PDU application frame arriving at an ingress junctor so as to extract therefrom the IP address of the destination, by searching in a cache table of the junctor for a pair (logical path, outbound direction) opposite the relevant IP address and opposite the ingress logical path and by using the translation obtained for all the cells of the PDU application frame, the cache table being updated by virtue of the routing information originating from the routing emulation function residing in the management module and in that it consists in transmitting a request to update the cache to the management module if the sought-after IP address is not located thereat or if the information opposite this address is too old.

Ins A5

Other characteristics and advantages of the invention will become apparent with the aid of the description which follows with reference to the appended drawings which represent:

- Figure 1a, a basic diagram of an ATM switch according to the prior art,

- Figure 1b, a basic diagram of an ATM switch with distributed architecture according to the prior art,

- Figures 2a to 2f, diagrams illustrating communication modes between users of an ATM network,

- Figure 3a, an example of steering an ATM cell through a switch during a point-to-multipoint connection,

- Figure 3b, an example of superimposing point-to-point or point-to-multipoint connections in an emulated LAN architecture,

- Figure 3c, a basic diagram of a routing between ELANs,

- Figure 3d and 3e, an illustration of the dynamic short-circuiting procedure implemented by the invention,

- Figure 4, an example of organizing an ATM switch for implementing the process according to the invention.

The process according to the invention makes it possible to alleviate the prior art drawback cited above as regards the excessive load of the management unit within which a virtual router of IP frames is embedded. The process makes it possible, inside an ATM switch, to achieve a genuine decentralization of the IP relay function (or IP forwarding) by limiting the role of the router to its function for calculating routes, which is already known from the prior art.

Figure 3c shows a case where this process is usable. The virtual router possesses as many router LEC modules as ELANs which it knows. If the user UA belonging to the ELAN A wishes to despatch an IP frame destined for a user UB, he begins by using the means for broadcasting on the ELAN A (BUS broadcast server). If the internal router in the switch knows of the existence of user UB, the LEC A module of this router, associated with the ELAN A, declares itself to the destination of all the IP frames destined for the user UB. Subsequently the user UA establishes and uses his direct ATM connection with the LEC A, according to the usual process specified in the LANE standard, so as to send frames destined for the user UB. According to the prior art, the frames have to backtrack to the internal router of the management module, so as to relay them to the user UB using the direct ATM connection which exists between the LEC B module and the user UB.

The process according to the invention specifies that for any application frame PDU arriving at a direct connection involving the LEC A module, the ingress junctor examines its first cell and extracts the IP address of the destination therefrom. It then runs through a cache table updated by virtue of the routing information originating from the management module, and finds therein opposite the IP address and opposite the inbound logical path a pair (logical path, outbound direction). The outbound direction is the identifier of the cross-connector junction involved in the direct connection between the LEC B module of the user UB. The logical path is an internal index making it possible to retrieve the logical path of this connection by virtue of an egress translation mechanism which will be described hereinbelow. If the ingress junctor does not find the IP address searched for in the cache table, it despatches a cache update request to the management module. The information found in the table then serves in translating the ATM header of each cell of the relevant PDU frame. This makes it possible, through a dynamic translation procedure, the translation table being modified potentially during the passage of each PDU frame, to thus establish a "dynamic short-circuit" between two point-to-point connections as shown diagrammatically in Figure 3d.

If the logical path found in the ingress translation table were simply the logical path associated with the direct connection between the LEC B module and the user UB, this would result in an interleaving between the various PDUs despatched simultaneously by various users of the ELANs concerned to the same user UB of ELAN B.

In order to avoid this drawback the process according to the invention makes provision for a double translation. To do this, the ingress and egress junctors of the switch are furnished in the manner represented in Figure 4, where the elements counterpart to Figure 3c are represented with the same references, with translation tables 9. In Figure 4, where only two ingress junctors 7i and 7k and only one egress junctor 7j are represented, the translation tables of the ingress junctors 7i and 7k bear the references 9i and 9k respectively and the translation table of the egress junctor 7j bears the reference 9j. These translation tables make it possible in the example represented to connect sending users UA_1, UA_2 to destination users UB and UC, one of whom UB features a local area network. Each cell originating from a sending user addresses a translation table 9 via a pair of values formed of a logical path number and of the IP address ($@IP_1$, $@IP_2$, etc.) of the destination user. The logical path and IP address pair is transformed by the translation table 9 into a pair of values composed of an index value VM and of an identifier number L_j for an egress junctor j involved in the direct connection between the LEC B module of the management unit 4 and the LEC UB of the destination user. In the example of Figure 4, the translation table 9i of the junctor 7i carries out the translation of the pair (VLi (UA_1), $@IP_1$) into a pair (VM (UA_1, UB), L_j) where VLi (UA_1) is the logical path associated in the junctor 7i with the direct connection between the user UA_1 and the LEC A module of the management module 4, $@IP_1$ is the IP address of the destination user belonging to the local area network UB, VM (UX, UY) is a connection internal index number allocated to each pair of users (UX, UY) and L_j is the identifier of the cross-connector junction 7j involved in the direct connection between the LEC B module and the destination user UB.

Likewise according to this same principle, the example of Figure 4 shows an example of a connection between a user UA_2 and two destination users having respective addresses $@IP_2$ and $@IP_3$. In this example, communications from UA_2 to UB and from UA_2 to UC are each carried out in the junctor 7k via the following respective translations:

$VL_k\ (UA_2,\ @IP_2) \rightarrow (VM\ (UA_2,\ UB),\ L_j$

and $VL_k\ (UA_2,\ @IP_3) \rightarrow (VM\ (UA_2,\ UC),\ L_j,$

According to the invention, each egress junctor 7j has available a large number of queues 11_n such that one of these queues, Fj (UX, UY), can be allocated one-to-one to each pair (UX, UY) where UY is a user whose direct connection between himself and his LEC router LEC Y passes through the junctor 7j. An egress translation table 9j arranged in each of the egress junctors 7j carries out an egress translation of the value of the index VM (UX, UY) into a pair (VLj (UY), Fj (UX, UY), where VLj (UY) is the logical path associated in the egress junctor 7j with the direct connection between the user UY and the LEC Y module and Fj (UX, UY) is the number of the queue of the junctor 7j allocated to the pair (UX, UY).

The internal index VM (UX, UY) must allow the egress translation on the junctor 7j relevant to the user UY. It is not therefore necessary for the function VM associating an index with each pair (UX, UY) to be one-to-one (injective) since the translation of the index is carried out within the context of junctor j. The number of indices required in the entire switch is therefore the maximum of the number of egress queues in each of the junctors. Moreover, neither is it necessary for the function VM to be defined for every pair (UX, UY) since two unspecified users UX and UY do not always need to converse, or they may sometimes converse without going via the router if they belong to the same ELAN. The allocating of the indices and of the egress queues can therefore be done dynamically, as a function of the needs expressed, for example in conjunction with the updating of the ingress translation caches.

Finally, an egress arbitrator which is within the scope of the person skilled in the art, and therefore not represented, carries out the extraction "in PDU mode" of the cells from the queues and their transmission over the physical interface. The "PDU mode" operation signifies that a queue is regarded as being ready to send only when it contains at least one complete PDU frame and when the arbitrator extracts only complete PDU frames.